Report on BooksOnline'11: 4th Workshop on Online Books, Complementary Social Media, and Crowdsourcing

Gabriella Kazai
Microsoft Research Cambridge
v-gabkaz@microsoft.com

Carsten Eickhoff
Delft University of Technology
c.eickhoff@tudelft.nl

Peter Brusilovsky
The University of Pittsburgh
peterb@pitt.edu

April 5, 2012

Abstract

The BooksOnline Workshop series aims to foster the discussion and exchange of research ideas and initiatives addressing challenges and exploring opportunities around large collections of digital or digitized books and complementary media. The fourth workshop in the series, BooksOnline'11[1] called for special attention to the role of social media and the phenomena of crowdsourcing in the context of online books, which are expected to be key in defining new user experiences in digital libraries and on the Web. The workshop boasted a high quality program, including keynote addresses by Ville Miettinnen, CEO of Microtask and Adam Farquhar, Head of Digital Library Technology at The British Library. From the accepted papers, two main themes became salient: 1) The role of relationships among authors, communities and books, and 2) Reading experiences and behaviours. This paper provides a summary of the workshop, its accepted contributions and the subsequent plenary discussion.

1 Introduction

In recent years, the volume of digitally available books has increased dramatically through electronic publishing and the digitization of physical books, e.g., Google Books Library Project, Project Gutenberg[2], the Million Book project[3], and the Open Content Alliance[4]. Large-scale digital libraries hold significant value to humanity by preserving knowledge and

[1] http://research.microsoft.com/booksonline11/
[2] http://www.gutenberg.org
[3] http://www.ulib.org
[4] http://www.opencontentalliance.org

making it widely accessible. Furthermore, such collections show great potential for cross-media integration and industrial exploration. At the same time, eBooks and eReaders are gaining wide acceptance and popularity. This is paralleled with social networking and content sharing platforms that accommodate millions of users who often dedicate great efforts to data collection, creation, annotation and verification. Harnessing such considerable forces has the potential to revolutionize the digital library sector as well as electronic reading, both technically and in terms of interaction paradigms.

To match the great momentum in creating on-line book repositories, the BooksOnline workshop series aims to foster research initiatives that are focused on innovation opportunities and challenges created by large collections of digital books and complementary media. Over the past four years, the workshop series covered a wide array of central topics. BooksOnline'08 [6] explored *enriched digital collections*, *usage scenarios* and *user experience* as well as *content representation* and *discovery services*. The second edition of the series, BooksOnline'09[5], concentrated on *design and interaction models* for digital libraries and the *search and evaluation* aspects. BooksOnline'10 [7] addressed *user aspects in design* and *infrastructure issues*.

This year, the focus of the workshop resided on online collaboration initiatives, either motivated by intrinsic common goals of an online community or by alternative incentives as demonstrated by the advances into commercial crowdsourcing. Several accepted papers targeted incorporating the wisdom of the crowd into the domain of electronic reading and the curation of digital book collections.

Since 2009, BooksOnline has been fortunate to receive sponsorship from Microsoft Research to reward outstanding scientific contributions and to grant seed funding for selected project proposals. In 2009, seed funding was awarded to George Buchanan (City University London, UK) "Building a Testbed for Document Annotation and Search" and Monica Landoni (University of Lugano, Switzerland) "Building a Bookshelf for Children" In 2010, Xavier Snelgrove and Ronald Baecker (University of Toronto, Canada) "A System for the Collaborative Reading of Digital Books with the Partially Sighted" and Claudia Hauff and Dolf Trieschnigg (University of Twente, The Netherlands) "Enhancing Access to Classic Children's Literature", received funding. In 2011, Marc-Allen Cartright, Henry A. Field and James Allan (University of Massachusetts) received the Best Paper award for their contribution "Evidence Finding using a Collection of Books" [2]. The BooksOnline'11 Best Project Seed Fund award is to be announced at the end of April. In this report, we summarize the BooksOnline'11 workshop, its contributions and achievements as well as key topics that emerged in the adjunct plenary discussion of all participants.

2 Workshop Program

In this section, we provide a brief summary of the workshop program, including identified challenges and key directions of the ensuing discussions.

2.1 Keynote by Adam Farquhar from The British Library

The workshop was opened by a highly-appreciated keynote address by Adam Farquhar (The British Library). As Head of Digital Library Technology, Adam gave an overview of the

[5]http://research.microsoft.com/en-us/um/cambridge/events/booksonline09/

huge scale at which large modern libraries such as The British Library operate. More than 3000 visitors are served on a daily basis from a book collection that spans more than 800 kilometers of shelves. Out of this vast collection, approximately 1% is digitized.

The emphasis of Adam's talk resided on a perception change that starts to see digitized books as more than mere digital copies of their paper equivalents. As a consequence of this changing mindset, numerous new challenges and opportunities surfaced. For example, richer-than-paper editions provide added value in the form of interactive molecular models of chemical compounds, multimedia content or a social dimension, e.g., in the form of collaborative passage highlighting across users.

Especially historic books that are of scholarly interest hold much more information than the mere textual and illustrative content. Different paper and vellum types, preparation techniques, bindings and page scrapings can often not be perceived by the eye or are lost in the scanning process. To preserve these aspects of the original, it is necessary to investigate dimensions such as different page-turning behaviour and optional higher-quality scans to emulate the properties of the paper version.

Adam closed his talk by proposing two open questions: (1) Matters of scale: How to adequately deal with the prospect of one day having indexed the entire collection? Current paradigms such as the offering of APIs for external use may not be practical any more and may instead require the code to be sent to the library for local execution. (2) Does all this technology make people read more? Are there books that deserve to remain dusty with good reason?

2.2 Tools, Communities and Crowds

Tim Reagan from Microsoft Research Cambridge was unfortunately not able to attend the workshop and present his accepted contribution "Tools for Whom: Readers, Fans, or Authors?" [10]. In his article, he discusses the possibility of book text visualization for different user groups. Based on Pullman's popular trilogy "His Dark Materials", he demonstrates how character appearances and mentions can serve different purposes for authors and publishing agents who plan and analyse their story lines and for readers who want to closely follow the character development. Interviewing users from those different groups, he finds that especially the fan community was positive about the tool, while Pullman himself was worried about such technology drastically influencing and potentially hindering the creative process.

Christoph Becker from Vienna University of Technology presented his paper "Quality Assurance in Document Conversion: A HIT?" [1]. He discusses the question of how to asses document identity across different formats and versions. The conversion of documents from one format to another or across media, e.g., by scanning textual sources, often introduces changes to the material's layout. Assessing the conversion quality in terms of proximity to the original can therefore prove to be crucial. To supplement and establish automatic quality measures, he describes a crowdsourcing-based method of rating conversion quality. One particular finding was the workers' difficulty in consistently agreeing on the definition of abstract concepts such as document identity. He proposed to, in the future, break down document comparison tasks into a number of element comparisons rather than comparing on document level.

2.3 Reading with a Purpose

Marc-Allen Cartright from The University of Massachusetts presented the BooksOnline'11 best paper contribution "Evidence Finding using a Collection of Books" [2] which was co-authored together with Henry A. Field and James Allan. He introduced the novel task of of Evidence Finding (EF). Its goal is finding confirming or refutative evidence for natural language assertions. The authors highlight the differences between the introduced problem and related fields such as Question Answering (QA). The approach taken and evaluated in their work phrases search engine queries based on the original assertions. These queries are subsequently issued to an index of 50000 digitally scanned books. A preliminary evaluation reports superior performance to straight-forward bag-of-words and Sequence Dependence models. Finding Evidence for composite assertions was identified as a hard problem that may require splitting into constituent facts and faceted presentation of facts.

Hélène de Ribaupierre from the University of Geneva presented the article "New Trends for Reading Scientific Documents" [4], joint work together with Gilles Falquet. She describes a user survey concerning the reading behaviour of over 90 research professionals. She argues that document retrieval would benefit from a more facet-oriented indexing to enable precise querying for desired information. Furthermore, such facets should be designed to support the natural tasks that motivate reading. I.e., different professional information needs may require the inspection of different document facets to be satisfied. The assumption is supported by the finding that the professional reading methodology differs significantly between researchers from different fields.

Heimo Müller from the Medical University of Graz presented "How to Carry Over Historic Books into Social Networks" [9], co-authored with Hermann A. Maurer. He defines a quality evaluation framework for ebooks and investigates several aspects of scan and OCR quality as well as identifying desirable interactive features. The described system, IIB (Interactive Internet Book), is an interesting approach to presenting historical books on-line as it aims to preserve the reading experience of real historic books while making them available to the broad audiences of online communities. Particular focus is put on interweaving the digital books with existing knowledge spaces such as Forums or Wikis to embed original historic information in digital information-sharing environments.

2.4 Authors, Links and Relations

Jaap Kamps from the University of Amsterdam talked about "The Impact of Author Ranking in a Library Catalogue" [5]. He proposes the use of expert finding methods in book retrieval in order to include an author score into the ranking function for book retrieval. He makes the noteworthy finding that author scores estimated based on an aggregate of each author's individual book scores can result in superior performance when directly compared to book scores. In this way, he demonstrates the merit of enriching existing library catalogue information by external or aggregate information sources. A particular use case for the inclusion of author scores is seen in novice researchers who enter a new field and are unaware of the distribution of expertise across authors.

Young-Min Kim from the University of Avignon presented "Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs" [8], a joint publication together with Patrice Bellot, Elodie Faath and Marin Dacos. She presents an automatic extraction scheme for bibliographic references from scientific literature in the hu-

manities. On a sizable corpus, they employ Conditional Random Fields (CRF) to extract reference fields such as author names or publication titles with convincing precision. Such information, in turn, can be used to enhance accurate resource retrieval at large scale.

James Allan from The University of Massachusetts presented "Mining Relational Structure from Millions of Books " [11], co-authored together with David A. Smith and R. Manmatha. The authors of this position paper propose to automatically extract the relational structure between OCR-scanned books by employing means of partial duplicate detection. Given a large-scale corpus of books, their method is assumed to enhance literature research as well as language analysis and authorship detection. An overview of preliminary results promises potential gains in effectiveness and efficiency.

2.5 Keynote by Ville Miettinnen from Microtask

The afternoon keynote was given by Ville Miettinnen (Microtask). In his talk, Ville gave an elaborate description of crowdsourcing both in terms of an altruistic social online phenomenon as well as a market. He introduced the history and key concepts involved in the creation of *Human Intelligence Tasks* (HIT). His particular emphasis resided on creative crowdsourcing efforts in which design tasks, song writing or even the creation of poems are outsourced to the crowd at professional result quality despite affordable pay rates.

After this general overview, Ville moved on to presenting the most recent and very successful project of his company Microtask. Microtask was involved in the digitization of the newspaper archives of the Finnish National Library. A particular challenge of the process was to resolve optical character recognition (OCR) errors. The Digitalkoot system addresses this challenge by forwarding ambiguous scan portions to an online game in which the players were asked to retype the displayed character sequences. Employing state-of-the-art insights into crowdsourcing mechanics, the system was able to reliably solve OCR errors without further costs as players were motivated by the prospect of helping to preserve their national cultural heritage.

An interesting observation that Ville shared in the closing of his talk was the fact that while the game was a huge success, several people approached him with the wish for a game-free, more professional interface, since they were only interested in the cultural preservation and not the entertainment aspect of the initiative.

2.6 Behind the Reading Experience

Adam Sofronijevic from the University of Belgrade introduced "Changes in Reading Research Proposition: Some Psychological Aspects of Reading 2.0" [12]. He contrasts conventional, solitary reading, in which the reader is isolated from other actors with the newly-introduced notion of *Reading 2.0*, a reading process that makes use of collaborative and social aspects, allowing the reader to interact with the author, the book content and other readers by means of digital technology.

Luca Colombo from the University of Lugano presented "Towards an Engaging e-Reading Experience" [3], a joint contribution together with Monica Landoni. He discusses the initial phase of a project for the development of engaging, immersive children's ebooks. The paper presents the outcome of a user study with elementary school children dedicated to their book preferences when reading for fun. The in-line integration of entertaining multimedia

content such as pieces of music or videos is foreseen to make reading more engaging for young audiences.

2.7 Bring-a-Challenge Discussion

In the course of the workshop day, all participants were encouraged to contribute concrete challenges from their domain to serve as a basis of a concluding plenary discussion.

The first major discussion item was dedicated to the fundamental question of purpose. What do people want ebooks for and what do they require from an ebook? The general consensus was that, at the moment, electronic books do not go far beyond being electronic ink, a direct digital representation of the paper original. However, substantial sensation can reside in the device rather than the book content itself. Besides the aspect of storage and portability, electronic reading devices could enhance reading by resource interlinking and content augmentation, both editorial and user generated. As a consequence of the new medium, reading is less location-dependent than previously as light-weight devices allow for the easy transportation of digital equivalents of otherwise heavy books. Furthermore, some rare or old books may not even be portable at all in their original forms. Geographically-aware ebooks may incorporate knowledge of the reader's current position into the search, augmentation and social aspects of ebooks. Finally, a key question remains, who searches within books? Digitization makes huge volumes easily searchable. It is not clear, however, who would be the beneficiaries of such technology, other than scholars. An interesting parallel was drawn between historic reference books and modern blogs. Such resources are collections of pointers to external material, compiled, curated and commented on by an editor whose expertise and credibility the reader relies on.

The second fundamental discussion topic was concerned with a future perspective on digitization. What happens when all books are digitized? Ambitious as this aim sounds, a state of nearly-exhaustive global indexing of popular library content can be expected in the coming 20-30 years, assuming today's digitization rates. Looking at this future vision, the participants identified resource interlinking as one of the next big challenges. Scholars across many areas would benefit from a topology of books, not simply connected by references but also dynamically groupable by semantic facets such as themes and topics, mentions of persons or concepts, their affiliation to certain historic periods and styles, etc. This would require rich annotation or sophisticated information mining techniques. Given such a ubiquitous library of interlinked resources, we can furthermore expect effects on the way and rate at which patents are granted or challenged, as novelty and ingenuity of contributions are assumed to be more easily determined.

3 Conclusion & Outlook

This year's edition of the BooksOnline workshop series focused on the role of users, social groups and crowds for addressing central tasks in the creation, annotation and maintenance of large digital book repositories. The ensuing discussion showed a fundamental need for better understanding of user behaviour in order to shape the ebook of the future that goes beyond being digital mirror images of paper books. Future editions of the workshop will aim to deepen the insights into this complex relationship by establishing and confirming use cases and interaction paradigms for future ebooks.

Acknowledgements We would like to thank ACM and CIKM for hosting the workshop, and the CIKM Workshops chair, Craig MacDonald, for his outstanding support in the organization.

We would also like to thank the members of the program committee for all their dedicated work in shaping the program: James Allan (University of Massachusetts Amherst, US), Nicholas Chen (University of Maryland, US), Bruce Croft (University of Massachusetts Amherst, US), Gilles Falquet (University of Geneva, Switzerland), Gene Golovchinsky (FX Palo Alto Laboratory, Inc., US), Ananda Gunawardena (Carnegie Mellon University, US), Geneva Henry (Rice University, US), Ivan Koychev (Sofia University, Bulgaria), Monica Landoni (University of Lugano, Switzerland), Birger Larsen (Royal School of Library and Information Science, Denmark), Ray Larson (University of California, Berkeley, US), Michael E. Lesk (Rutgers University, US), Paolo Manghi (CNR, Italy), Riccardo Mazza (University of Lugano, Switzerland), Catherine C. Marshall (Microsoft Corporation, US), Natasa Milic-Frayling (Microsoft Research, UK), John Ockerbloom (University of Pennsylvania, US), Jerme Picault (Alcatel-Lucent Bell Labs), Prakash Reddy (Hewlett-Packard), Nina Wacholder (Rutgers University, US), Ian Witten (University of Waikato, New Zealand).

Special thanks are due to the paper authors, the invited speakers: Adam Farquhar and Ville Miettinnen, and all the participants for a lively workshop.

We would like to thank Microsoft Research for their sponsorship including the best paper and best project seed fund awards.

References

[1] Christoph Becker. Quality assurance in document conversion: a hit? In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 5–10, 2011.

[2] M.A. Cartright, H.A. Feild, and J. Allan. Evidence finding using a collection of books. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 11–18. ACM, 2011.

[3] Luca Colombo and Monica Landoni. Towards an engaging e-reading experience. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 61–66, 2011.

[4] Hélène de Ribaupierre and Gilles Falquet. New trends for reading scientific documents. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 19–24, 2011.

[5] Jaap Kamps. The impact of author ranking in a library catalogue. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 35–40, 2011.

[6] P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson. Booksonline08: Proceeding of the 2008 acm workshop on research advances in large digital book repositories. *ACM, New York*, 2008.

[7] G. Kazai and P. Brusilovsky. 3rd booksonline workshop: research advances in large digital book repositories and complementary media. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1967–1968. ACM, 2010.

[8] Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. Automatic annotation of bibliographical references in digital humanities books, articles and blogs. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 41–48, 2011.

[9] Heimo Müller and Hermann A. Maurer. How to carry over historic books into social networks. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 25–34, 2011.

[10] Tim Regan. Tools for Whom: Readers, Fans, or Authors? In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 1–4, 2011.

[11] David A. Smith, R. Manmatha, and James Allan. Mining relational structure from millions of books: position paper. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 49–54, 2011.

[12] Adam Sofronijevic. Changes in reading research proposition: some psychological aspects of reading 2.0. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 57–60, 2011.